Weekly Report

Period: 2016/7/25-2016/7/31
Reporter: Li Zongzhuang

Visual Exploration and Analysis of Knowledge Graph

Li Zongzhuang

Abstract: The Knowledge Graph is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources. The Knowledge Graph uses the power of semantics, and wants to improve search precision and effectiveness by building the semantic web. Visualization is the study of (interactive) visual representations of abstract data to reinforce human cognition. The visualization of Knowledge Graph can Effectively improve the efficiency of the user to complete the search target and precision. Data integration and correlation analysis reasoning is one of the best visual analysis applications. At this time, There are many applications have been exploited based on this concept.

1. Introduction

Knowledge graphs on the Web are a backbone of many information systems that require access to structured knowledge. The idea of feeding intelligent systems and agents with general, formalized knowledge of the world dates back to classic Artificial Intelligence research in the 1980s. Then, with the advent of Linked Open Data sources like DBpedia, and by Google's announcement of the Google Knowledge Graph in 2012, representations of general world knowledge as graphs draw a lot of attention again.

In 2014, Google announced a new initiative,called the Knowledge Vault, which derives much of its data from the Knowledge Graph and the sources thereof, as well as harvesting its own data, ranking its reliability and compiling all results into a database of over 1.6 billion facts collected by machine learning algorithms.

In a paper , author shows what is a knowledge graph:

1. mainly describes real world entities and their interrelations, organized in a graph.
2. defines possible classes and relations of entities in a schema.
3. allows for potentially interrelating arbitrary entities with each other.
4. covers various topical domains.

Vision is the most important channels to the information of the outside world. Visualization is the data technology of interactive visual expression. On the century of big data, The ability of processing data is far behind the ability to get the data. The amount of data contained in Knowledge Graph is huge, so the visualization can be an important means of Knowledge Graph data processing. It can help us find the phenomena and laws faster and achieve the goal. However,

the research about the visualization of Knowledge Graph is relatively shallow.

2. Knowledge Graph

2.1 Knowledge Graphs

There are many ways to build knowledge graphs. They can be curated like *Cyc*, edited by the crowd like *Freebase* and *Wikidata*, They can also be extracted from large-scale, semi-structured web knowledge bases such as Wikipedia, *DBpedia* and *YAGO*. Furthermore, information extraction methods for unstructured or semi-structured information are proposed, which lead to knowledge graphs like *NELL*, *PROSPERA*, or *KnowledgeVault*.

Freebase, a public, editable knowledge graph with schema templates for most kinds of possible entities. The last version of Freebase contains roughly 50 million entities and 3 billion facts. Freebase's schema comprises roughly 27,000 entity types and 38,000 relation types.[1] It have been shutdown because of company aquired by Google.

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. Keys are mapped to properties in that ontology. Based on those mappings, a knowledge graph can be extracted. It contains 6.2 million entities and 187 million statements about those entities. [2]The ontology comprises 735 classes and 2,800 relations.[]

After the shutdown of Freebase, the data contained in Freebase is subsequently moved to Wikidata.[]In Wikidata, for each axiom, it's provenance metadata can be included.[] Wikidata contains roughly 19 million instances and 100 million statements. [3]Its schema defines 23,000 types and 1,600 relations.

Google's Knowledge Graph was introduced to the public in 2012, and it was the tern knowledge graph being coined. Google's Knowledge Graph display was added to Google's search engine in 2012. Once a user search one thing, it provides structured and detailed information about the topic in addition to a list of links to other sites. According to Google, the information in the Knowledge Graph is derived from many sources, including the CIA World Factbook, Wikidata, and Wikipedia.It contains 18 billion statements about 570 million entities, with a schema of 1,500 entity types and 35,000 relation types.[]

Never-Ending Language Learning is an implementation of the Read the semi-structured Web data approach. [] As opposed to DBpedia, all facts recorded by NELL can be tracked according to its provenance and a degree of confidence.[] Nell2RDF platform can transform the data generated by NELL into state of the art Linked Data, following best practices.[] NELL has been learning to read the web 24 hours/day since January 2010, and so far has acquired a knowledge base with over 80 million confidence weighted beliefs (e.g., servedWith(tea, biscuits)).[]

---

[1] http://www.freebase.com.
[2] http://wiki.dbpedia.org/services-resources/datasets/dataset-2015-10/dataset-2015-10-statistics
[3] https://tools.wmflabs.org/wikidata-todo/stats.php

Name	Instances	Facts	Types	Relations
DBpedia (English)	4,806,150	176,043,129	735	2,813
YAGO	4,595,906	25,946,870	488,469	77
Freebase	49,947,845	3,041,722,635	26,507	37,781
Wikidata	15,602,060	65,993,797	23,157	1,673
NELL	2,006,896	432,845	285	425
OpenCyc	118,499	2,413,894	45,153	18,526
Google's Knowledge Graph	570,000,000	18,000,000,000	1,500	35,000
Google's Knowledge Vault	45,000,000	271,000,000	1,100	4,469
Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

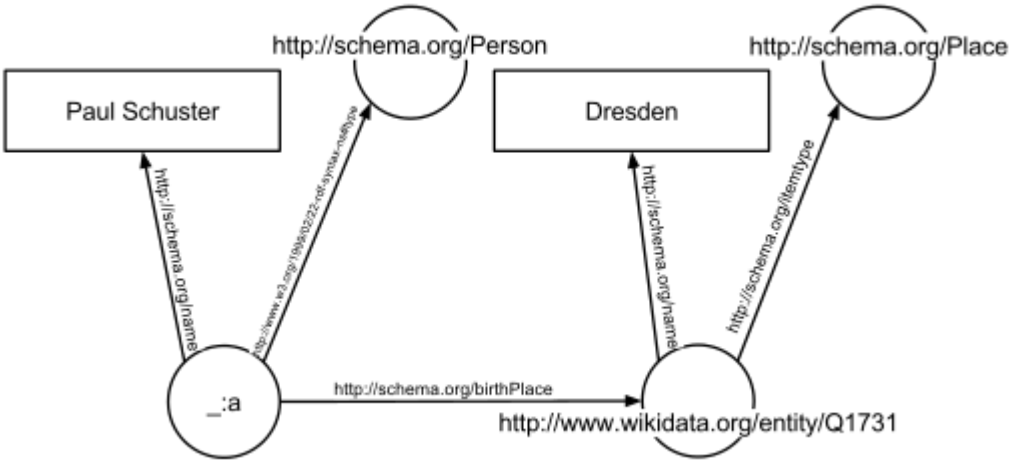
An overview about these knowledge graphs.

2.2 Semantic Web

The core concept of Knowledge is the introduction of the semantic, which means that let the computers know the semantic judgments. The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF).

The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users.

The concept of the Semantic Network Model was formed in the early 1960s by the cognitive scientist Allan M. Collins, linguist M. Ross Quillian and psychologist Elizabeth F. Loftus as a form to represent semantically structured knowledge. When applied in the context of the modern internet, it extends the network of hyperlinked human-readable web pages by inserting machine-readable metadata about pages and how they are related to each other. The Semantic Web is regarded as an integrator across different content, information applications and systems.



Semantic Web uses ontology because two databases may use different identifiers for what is in fact the same concept. Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches—the search program can look for only those pages that refer to a precise concept instead of all the ones using ambiguous keywords.

3. Visualization

3.1 Graph data visualization

Graph data is an important component in data. The visualizations about graph data are often

presented by node-link graph.

3.2 High-dimensional data visualization

There are many data sets have more than one dimension. So many visualization tools have been invented to present high-dimensional data. The results got by Knowledge Graph often have many properties. That means we can get some revelation.

3.3 Another types visualization

Because of the difference of goals, there can be many visualization schemes. Maybe we can learn more from them.

4. Application softwares

Based on the theory of human-computer interaction, there are a lot of software is put forward based on the semantic web. This kind of software focus on data integration and correlation analysis. Data integration made in background automatically, and data correlation analysis mainly rely on people's reasoning ability as well as front end some interactions. That's the best application, which give full play to the people the calculation of analytical reasoning skills and computer specialty.

There are many applications in this area, such palantir, IBM i2, Tableau and so on.

Palantir, IBM i2 which are designed aiming to dig the unstructured and network relations, analysis and reasoning, and stresses the participation of human intelligence is the best example of visual analysis. Most customers of them are on the safe, seeking field.

Firstly we introduce Palantir. Palantir software originated in 2004, Thiel investment Dr Carp to set up Palantir company to develop a large software based on ebay online transaction fraud to give priority to with intelligence analysis.

Until 2013, Palantir software has achieved great success.The CIA is its biggest customer, he helped the CIA track the whereabouts of Bin Laden in Afghanistan.

The Palantir platform architecture was designed to be extensible open source or plug-in type system, allows the user to add or remove modules.Palntir also provides the public API, call for outside developers. Its data format is designed as XML format, for users to customize and extend freedomly.

Palantir's core elementis that all things are established by using Ontology. Palantir introduces Palantir Dynamic Ontology Manager (the PDOM) concept, which is used in the fields related transactions based on Ontology modeling, operation, management and correlation, analysis, reasoning, and visualization. It has three features: highly free and nondestructive data abstraction, data format and API publicly.

Palantir does not define a fixed processing data formats. The original data according to the original form. Therefore, Palantir designed a is located in the abstract data Model, database and client POM (Palantir Object Model).

The basic concept of POM is an object. Object is an expression of all sorts of transaction data containers in the world. Includes: physical objects (people, places, organizations, or other name); Event object (happened at some point in time or interval); Document object (unstructured data, such as E-mail messages or news reports); Multimedia objects (rich media content such as video, audio, and images). An object has a name and unique ID, but does not contain data. In addition, the object also contains a number of confidentiality properties: feature, links, tags, etc. The reason

why these attributes are secret is that a record of each attribute has a data source, to track the attribute data source. At the same time, any attribute of an object can have multiple instances.

Ontology is the categorization of the objective world. Palantir dynamic ontology is to analysis the classification of the visual.

The dynamic ontology method is used in all aspects of the Palantir system, such as the user's query, the network analysis. Ontology method affects every aspect of users' interaction with the system, and therefore is very important to the efficiency of analysis.

The customer can adopt Palantir dynamic ontology tests, design, custom edit and manage all ontology. The specific function of it are: 1) adding new objects, attributes, and links to the dynamic ontology; 2) delete useless objects, attributes, and links; 3) change the function and behavior of objects, attributes, and links; 4) generate tag name when inputting data; 5) generate an alias, used in the system menu type lookup with fuzzy search function; 6) validators, used for specific attributes matching their conditions provide rules; 7) the comparator, used for fuzzy comparison.

Palantir system consists of five functional modules.

1. Data integration: support heterogeneous, multi-source, unstructured data integration.
2. Advanced search, find and analyze. Support association, temporal and geographical space, forecasting, statistics, behavior, and network analysis.
3. Knowledge management. Support the management of the knowledge which user have gained.
4. Collaboration: support remote, no Internet users and Web users collaborative analysis tasks, sharing news, subject object and analysis results.
5. Algorithm engine: support all kinds of data transformation algorithm, support the Pb level data parallel processing.

IBM ® i2 series product is a specially designed for investigation, analysis, investigators visualization data analysis software. It can change structured, semi-structured, and unstructured data into graphics. It provide the analyst with a straightforward entity relationship diagram, and provides a rich visual analysis algorithms and analysis tools. It helps analysts quickly find clues to solve crimes and valuable information, which improves work efficiency and helps identify, predict and prevent crime, terrorism and money laundering and fraud activities.

IBM i2 also offers a wide range of visual analysis algorithms and analysis tools, including link analysis, path analysis, cluster analysis, social network analysis, etc.

Here is a conclusion of some software about data.

	Open source	Product	Technique	Applications	Easy To Use	Efficiency	Support for unstructured data
R	Yes	No	Statistical analysis, and visualization	Numeric statistics	Optimal	Poor	Poor
Pentaho	Yes	Yes	Data integration, data mining	Business intelligence	Good	Good	Good

Spotfire	No	Yes	Visualization, analysis	Business intelligence, biological	Optimal	Optimal	Poor
MicroStrategy	No	Yes	Data analysis, database,OLAP	Business intelligence, The report	Optimal	Good	Poor
Tableau	No	Yes	Data visualization	Business intelligence, Office automation	Optimal	Optimal	Good
IBM i2	No	Yes	Data integration, correlation analysis and reasoning	Public safety	Optimal	Optimal	Optimal
Palantir	No	Yes	Data integration, correlation analysis and reasoning	Public security, commercial, financial, fraud	Optimal	Optimal	Optimal